AD-A080 911   WISCONSIN UNIV-MADISON DEPT OF STATISTICS   F/G 12/1
USE OF BOX-COX TRANSFORMATION WITH BINARY RESPONSE MODELS.(U)
AUG 79   V M GUERRERO, R A JOHNSON   N00014-78-C-0722

UNCLASSIFIED   UWIS-DS-79-575   NL

1 OF 1
AD
A080911

END
DATE
FILMED
3-80
DDC

(12) LEVEL II

ADA080911

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN

Madison, Wisconsin

(15) N00014-78-C-0722

(12) 1̶2̶

(14) UWIS-DS-79-575

DDC FILE COPY

(11) Aug 79

(9) TECHNICAL REP[redacted]T, [redacted]

August 1979

(6) USE OF BOX-COX TRANSFORMATION WITH BINARY RESPONSE MODELS.

by

(10) Victor M. Guerrero [redacted] Richard A. Johnson
University of Wisconsin

80 2 20 004

400243

slt

Use of the Box-Cox Transformation with Binary Response Models

by

Victor M. Guerrero
University of Wisconsin

Richard A. Johnson
University of Wisconsin

Summary

The power transformation suggested by Box and Cox (1964) is applied to the odds ratio to generalize the logistic model and to parameterize a certain type of lack of fit. Transformation of the design variable within the context of the dose-response problem is also considered.

1. Introduction

The use of linear logistic regression models is now widespread (c.f. Cox (1970), Nerlove and Press (1973)). By introducing an additional parameter that allows for other than logarithmic transformations of the odds-ratio, we extend their applicability. In the spirit of Box and Cox (1964), our transformations are data based. Viewed in another way, by determining plausible values for the transformation parameter, we are able to decide whether or not the logarithm is an appropriate transformation. In this sense, we obtain a single parameter lack of fit criterion for linear logistic models.

As suggested by Cox (1970), we also consider transformations of the independent variable in the context of the dose-response problem. In this case, we obtain the correct asymptotic covariance matrix of the estimators by allowing the assumed model to be incorrect.

2. Linear Models for Proportions

In the linear logistic regression model, the assumption is made that a linear relationship is appropriate for linking the log-odds ratio of the dependent variable to several explanatory variables. That is, if $Y_1,\dots,Y_n$ are independent 0-1 random variables with

$p_j = P\{Y_j = 1\}$, then

$$\log\left(\frac{p_j}{1-p_j}\right) = \underset{\sim}{\beta}'\underset{\sim}{X}_j, \qquad j = 1,\dots,n \qquad (1)$$

where $\underset{\sim}{\beta}$ is a q-dimensional vector of unknown parameters and $\underset{\sim}{X}_j = (1, X_{j1},\dots,X_{j(q-1)})'$ is the j^{th} vector of observations on (q-1) explanatory variables.

Alternative transformations of the probability p_j have been suggested for linearizing purposes. Among the most common alternatives are the integrated normal, the arc-sine and the identity transformations. However, the arc-sine and the identity have finite ranges which sometimes limits their usefulness. On the other hand, the choice between the logistic and the normal functions is usually a matter of taste although, in recent times, the logistic model seems to have more advocates (c.f. Nerlove and Press (1973)). Prentice (1976a) gives some reasons why the odds-ratio should be particularly considered in retrospective studies. Also, as pointed out by Cox (1970, p. 26), differences on a logistic scale have simpler interpretation in terms of the odds for success against failure.

The previous considerations have led us to study a "natural" extension of the linear logistic regression model. We assume that some power transformation of the odds-ratio satisfies a linear model. That is,

$$\left(\frac{p_j}{1-p_j}\right)^{(\lambda)} = \underset{\sim}{\beta}'\underset{\sim}{X}_j \qquad \text{for } j = 1,\dots,n \qquad (2)$$

where $\underset{\sim}{\beta}$ and $\underset{\sim}{X}_j$ are as in (1) and

$$\left(\frac{p_j}{1-p_j}\right)^{(\lambda)} = \begin{cases} \log\left(\frac{p_j}{1-p_j}\right) & \text{if } \lambda = 0 \\ \frac{1}{\lambda}\left[\left(\frac{p_j}{1-p_j}\right)^{\lambda} - 1\right] & \text{if } \lambda \neq 0 . \end{cases}$$

Model (2) focusses again on the odds ratio, includes as a special case the logistic model and can be used in all situations in which logistic regression is generally employed. One thing that should be remembered is that one extra degree of freedom will be used in estimating the transformation parameter λ.

In the remainder of this section, we will assume that there exist k different conditions at which successes are recorded. Let n_i be the number of observations at condition i and r_i the number of successes $(i = 1,\ldots,k)$. We tentatively assume that there is some value for λ_0 such that (2) holds. Then

$$p_i(\underset{\sim}{\theta}_0) = \begin{cases} [1 + \exp(-\underset{\sim}{\beta}_0' \underset{\sim}{X}_i)]^{-1} & \text{if } \lambda_0 = 0 \\ [1 + (1+\lambda_0 \underset{\sim}{\beta}_0' \underset{\sim}{X}_i)^{-1/\lambda_0}]^{-1} & \text{if } \lambda_0 \neq 0 \end{cases} \quad (3)$$

for some parameter values $\underset{\sim}{\theta}_0' = (\underset{\sim}{\beta}_0', \lambda_0)$.

The likelihood function for $n = \sum_{i=1}^{k} n_i$ observations is given by

$$L_n(\underset{\sim}{\theta}) = \prod_{i=1}^{k} \binom{n_i}{r_i} p_i^{r_i}(\underset{\sim}{\theta})[1 - p_i(\underset{\sim}{\theta})]^{n_i - r_i}$$

which yields the log-likelihood

$$\ell_n(\underset{\sim}{\theta}) = c + \sum_{i=1}^{k} \{r_i \log[p_i(\underset{\sim}{\theta})] + (n_i - r_i)\log[1 - p_i(\underset{\sim}{\theta})]\}$$

or

$$\ell_n(\underset{\sim}{\theta}) = \begin{cases} c + \sum_{i=1}^{k} r_i \underset{\sim}{\beta}' \underset{\sim}{X}_i - \sum_{i=1}^{k} n_i \log[1 + \exp(\underset{\sim}{\beta}' \underset{\sim}{X}_i)] & \text{if } \lambda = 0 \\ c + \sum_{i=1}^{k} \frac{r_i}{\lambda} \log(1 + \lambda \underset{\sim}{\beta}' \underset{\sim}{X}_i) - \sum_{i=1}^{k} n_i \log[1 + (1 + \lambda \underset{\sim}{\beta}' \underset{\sim}{X}_i)^{1/\lambda}] & \text{if } \lambda \neq 0 \end{cases} \quad (4)$$

with $c = \sum_{i=1}^{k} \{\log(n_i!) - \log(r_i!) - \log[(n_i - r_i)!]\}$.

Even when model (3) is not correct, we are able to establish the strong consistency of the MLE.

<u>Theorem 1</u>: Let $p_i(\underset{\sim}{\theta})$ be given by (3) and p_i be the unknown true probability of success. Suppose that

(i) the parameter space is a compact subset of R^{q+1}

(ii) $\lim_{n\to\infty} \frac{n_i}{n} = s_i$, with $s_i \in (0,1)$ and $\sum_{i=1}^{k} s_i = 1$

(iii) $H(\theta) = \sum_{i=1}^{k} s_i \{p_i \log\left[\frac{p_i(\underset{\sim}{\theta})}{p_i}\right] + (1-p_i)\log\left[\frac{1-p_i(\underset{\sim}{\theta})}{1-p_i}\right]\}$ has a unique global maximum at $\underset{\sim}{\theta} = \underset{\sim}{\theta}_0$.

Then, $\lim_{n\to\infty} \hat{\underset{\sim}{\theta}}_n = \underset{\sim}{\theta}_0$ with probability one.

Proof: Let us write $\hat{p}_i = \frac{r_i}{n_i} = \sum_{j=1}^{n_i} Y_{ij}/n_i$. Then, almost surely $\hat{p}_i \to p_i$ as $n_i \to \infty$, for $i = 1,\ldots,k$. Applying Stirling's formula for factorials, we have

$$\log(n_i!) - \log(r_i!) - \log[(n_i - r_i)!] = -n_i[\hat{p}_i \log(\hat{p}_i) + (1-\hat{p}_i)\log(1-\hat{p}_i)] + O(n_i^{-1})$$

on an almost sure set, where $O(n_i^{-1})$ is uniform in $\underset{\sim}{\theta}$. So, it follows that

$$\frac{1}{n}\ell_n(\underset{\sim}{\theta}) = \sum_{i=1}^{k} \frac{n_i}{n}\{\hat{p}_i \log[p_i(\underset{\sim}{\theta})] + (1-\hat{p}_i)\log[1-p_i(\underset{\sim}{\theta})] - \hat{p}_i \log(\hat{p}_i) - (1-\hat{p}_i)\log(1-\hat{p}_i)\} + o(1) \quad \text{as} \quad n \to \infty \tag{5}$$

with probability one.

Now, from (4), $\ell_n(\underset{\sim}{\theta})$ is seen to be continuous in $\underset{\sim}{\beta}$ and λ. By compactness of the parameter space and continuity of $\ell_n(\underset{\sim}{\theta})$ we obtain, with probability one

$$\lim_{n\to\infty} \frac{1}{n}\ell_n(\underset{\sim}{\theta}) = \sum_{i=1}^{k} s_i\left(p_i \log\left[\frac{p_i(\underset{\sim}{\theta})}{p_i}\right] + (1-p_i)\log\left[\frac{1-p_i(\underset{\sim}{\theta})}{1-p_i}\right]\right)$$

uniformly in $\underset{\sim}{\theta}$. Because the limit has a maximum at $\underset{\sim}{\theta}_0$, $\hat{\underset{\sim}{\theta}}_n \xrightarrow{a.s.} \underset{\sim}{\theta}_0$ as $n \to \infty$. ∎

Remark: For each i, the Kullback-Leibler information number between the true probability distribution and model (3) is

$$E_{Y_i}\left\{\log\left[\frac{P[Y_i = y]}{p_i(\underset{\sim}{\theta})}\right]\right\} = p_i \log\left[\frac{p_i}{p_i(\underset{\sim}{\theta})}\right] + (1-p_i)\log\left[\frac{1-p_i}{1-p_i(\underset{\sim}{\theta})}\right].$$

Then, we notice that

$$\lim_{n\to\infty} \frac{1}{n}\ell_n(\underset{\sim}{\theta}) = -\sum_{i=1}^{k} s_i\, E_{Y_i}\left\{\log\left[\frac{P[Y_i = y]}{p_i(\underset{\sim}{\theta})}\right]\right\};$$

Thus, maximizing the log-likelihood under model (3) is asymptotically equivalent to minimizing the Kullback-Leibler information number between the true and the proposed models.

The asymptotic normality of $\hat{\underset{\sim}{\theta}}_n$ is stated in the following theorem.

Theorem 2: If (3) holds for some $\underset{\sim}{\theta}_0' = (\underset{\sim}{\beta}_0', \lambda_0)$ $\sqrt{n}(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0) \xrightarrow{d} N_{q+1}(\underset{\sim}{0}, V)$ as $n_i/n \to s_i$ $(0 < s_i < 1)$. Where

$$V^{-1} = -\nabla^2\left(\sum_{i=1}^{k} s_i\left\{p_i \log\left[\frac{p_i(\underset{\sim}{\theta})}{p_i}\right] + (1-p_i)\log\left[\frac{1-p_i(\underset{\sim}{\theta})}{1-p_i}\right]\right\}\right).$$ ∎

We suggest to obtain the MLE's by solving the likelihood equations for $\underset{\sim}{\beta}$, at a fixed value of λ and then varying λ until the log-likelihood is maximized. The normal equations for $\beta_1,\ldots,\beta_q$ are

$$\sum_{i=1}^{k} X_{iu}(1+\lambda\underset{\sim}{\beta}'\underset{\sim}{X}_i)^{-1}[r_i - n_i p_i(\underset{\sim}{\theta})] = 0, \quad u = 1,\ldots,q.$$

A consistent estimator of the variance-covariance matrix of $\hat{\underset{\sim}{\theta}}_n$ can be

obtained by inverting $\left(-\frac{\partial^2 \ell_n(\theta)}{\partial\theta_u \partial\theta_v}\right)_{(q+1)\times(q+1)}$ and evaluating it at $\underset{\sim}{\theta} = \hat{\underset{\sim}{\theta}}_n$, where

$$\frac{\partial^2 \ell_n(\theta)}{\partial\beta_v \partial\beta_u} = \sum_{i=1}^{k} X_{iu} X_{iv} (1+\lambda \underset{\sim}{\beta}' \underset{\sim}{X}_i)^{-2} \{\lambda[n_i p_i(\underset{\sim}{\theta}) - s_i] - n_i p_i(\underset{\sim}{\theta})[1-p_i(\underset{\sim}{\theta})]\}, \qquad u,v = 1,\ldots,q$$

$$\frac{\partial^2 \ell_n(\theta)}{\partial\lambda\, \partial\beta_u} = \sum_{i=1}^{k} X_{iu} (1+\lambda \underset{\sim}{\beta}' \underset{\sim}{X}_i)^{-2} \{\underset{\sim}{\beta}' \underset{\sim}{X}_i [n_i p_i(\underset{\sim}{\theta}) - s_i] - n_i \lambda^{-2} [\lambda \underset{\sim}{\beta}' \underset{\sim}{X}_i - (1+\lambda \underset{\sim}{\beta}' \underset{\sim}{X}_i)\log(1+\lambda \underset{\sim}{\beta}' \underset{\sim}{X}_i)] p_i(\underset{\sim}{\theta}) \cdot [1-p_i(\underset{\sim}{\theta})]\}, \qquad \lambda \neq 0, \quad u = 1,\ldots,q$$

$$\frac{\partial^2 \ell_n(\theta)}{\partial\lambda^2} = \sum_{i=1}^{k} \lambda^{-2} \{[2\underset{\sim}{\beta}' \underset{\sim}{X}_i (1+\lambda \underset{\sim}{\beta}' \underset{\sim}{X}_i)^{-1} - 2\lambda^{-1} \log(1+\lambda \underset{\sim}{\beta}' \underset{\sim}{X}_i) + \lambda(\underset{\sim}{\beta}' \underset{\sim}{X}_i)^2 (1+\lambda \underset{\sim}{\beta}' \underset{\sim}{X}_i)^{-2}][n_i p_i(\underset{\sim}{\theta}) - s_i] - n_i [\underset{\sim}{\beta}' \underset{\sim}{X}_i (1+\lambda \underset{\sim}{\beta}' \underset{\sim}{X}_i)^{-1} - \lambda^{-1} \log(1+\lambda \underset{\sim}{\beta}' \underset{\sim}{X}_i)]^2 p_i(\underset{\sim}{\theta}) \cdot [1-p_i(\underset{\sim}{\theta})]\}, \qquad \lambda \neq 0 .$$

3. A 2×2^4 Factorial Arrangement

The situation we are considering covers the 2×2^m factorial system which arises frequently in either designed experiments or survey

studies, when the measurement of interest is the proportion of units with certain characteristics. The example that follows may be considered "classical" in the sense that many authors writing on categorical data have studied it (c.f. Cox (1970), Fienberg (1977)). Dyke and Patterson (1952) were the first to perform a maximum likelihood analysis on data collected by Lombard and Doering. Four factors were considered important in affecting the probability of getting a good score in a test on cancer knowledge. The factors being: (a) newspapers, (b) radio, (c) solid reading, and (d) lectures.

In the study, the aim was to estimate the main effects. Thus, we considered a model containing only six parameters, namely λ, β_1 (overall mean), β_2 (newspapers), β_3 (radio), β_4 (solid reading) and β_5 (lectures). The MLE of $\underset{\sim}{\theta}$ was obtained by maximizing the log-likelihood first with respect to $\underset{\sim}{\beta}$, for fixed λ, and then searching for a maximum over the values of λ. Figure 1 shows the graph of maximized log-likelihood function. The value of $\hat{\lambda}$, as read off the graph is .425, with a 95% confidence interval from -.112 to 1.104. Even though the value of $\hat{\lambda}$ is not significantly different from zero, we perform the analysis on this new scale. The original data and results of the analysis are presented in Table 1.

It should be noticed that $\hat{\lambda}$ is selected in accordance to the model proposed. Thus $\hat{\lambda}$ obtained under a model which takes into account only main effects, is not necessarily the best scale for a model which also contains interactions. To illustrate this point we introduce the first order interactions of lectures with the three

Figure 1
Maximized Log-likelihood Function

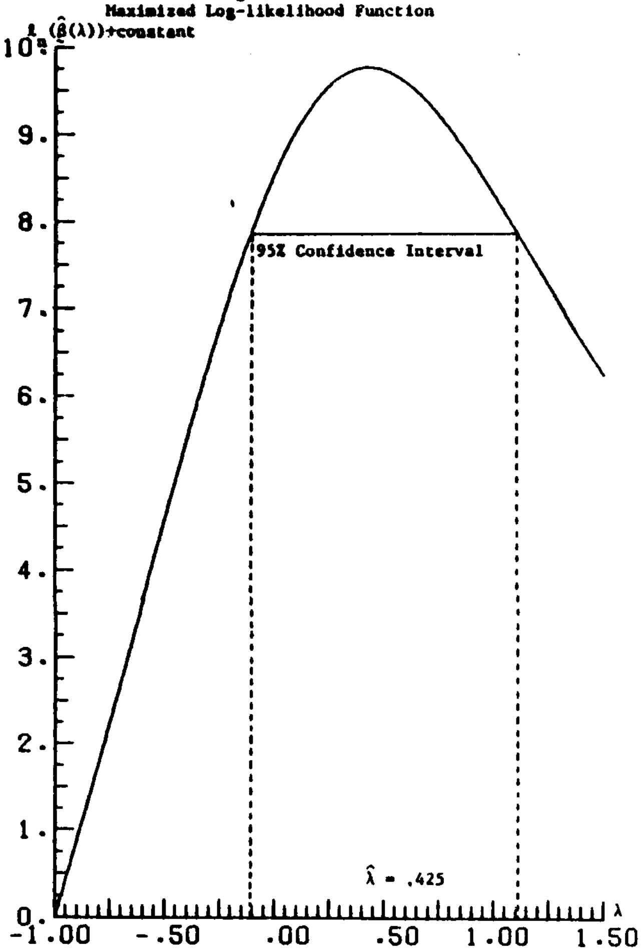

Table 1

CLASSIFICATION OF INDIVIDUALS WITH RESPECT TO CANCER KNOWLEDGE
(Taken from Dyke and Patterson (1952))

Factor Combination	No. of Trials	No. with Good Scores	Observed Proportion	Estimated Proportion (λ = .425)	Contribution to Chi-Square
1	477	84	.176	.174	.010
a	231	75	.325	.321	.007
b	63	13	.206	.244	.277
c	150	67	.447	.415	.206
d	12	2	.167	.292	.459
ab	94	35	.372	.392	.058
ac	378	201	.532	.543	.043
ad	13	7	.538	.437	.171
bc	32	16	.500	.481	.013
bd	7	4	.571	.364	.523
cd	11	3	.273	.521	.623
abc	169	102	.604	.596	.007
abd	12	8	.667	.501	.330
acd	45	27	.600	.627	.020
bcd	4	1	.250	.576	.313
abcd	31	23	.742	.670	.080
Total	1729	668	—	—	3.137

other factors, as in Dyke and Patterson (1952). The estimates obtained under the two models using the same $\hat{\lambda} = .425$ are presented in Table 2.

Table 2

MLE'S UNDER TWO DIFFERENT MODELS

	Model 1 (Main Effects Only)	Model 2 (Main Effects and Interactions with Lectures)
Mean	$-.1577 \pm .0928$	$-.1428 \pm .1120$
Newspapers (a)	$.2492 \pm .0395$	$.4410 \pm .1089$
Radio (b)	$.1207 \pm .0565$	$.2565 \pm .1080$
Solid Reading (c)	$.4106 \pm .0404$	$.2619 \pm .1089$
Lectures (d)	$.2009 \pm .0966$	$.2259 \pm .1120$
(ab)	-	$.2086 \pm .1089$
(bd)	-	$.1601 \pm .1080$
(cd)	-	$.1694 \pm .1089$
χ^2	3.137, 10 d.f.	.838, 8 d.f.

Thus, we observe that none of the interactions is significant at the 5% level. However, in a different scale, namely $\lambda = 0$, Dyke and Patterson found the interaction (cd) to be significant.

4. Transformation of the Design Variable and the Dose-Response Problem

The general dose-response problem that will be studied occurs when k groups of subjects are put under experiment. Corresponding to each of the k groups, there is a particular dosage level to be tested. Let us suppose that r_i responses are obtained when n_i subjects are studied at dosage x_i, for $i = 1,\ldots,k$. The problem then is to fit a cumulative probability distribution to the observed sigmoid response curve. The probability of observing r_i responses at dosage level x_i is

$$P(r_i|x_i) = \binom{n_i}{r_i} p_i^{r_i} (1-p_i)^{n_i - r_i}$$

where $p_i = P\{Y=1|x_i\} = P\{T<x_i\} = G(x_i)$ is the probability of an individual response $(Y = 1)$. Further, it is assumed that p_i can be expressed in terms of a tolerance distribution G associated with T.

It is generally understood that a symmetric tolerance distribution will adequately describe the data if a logarithmic transformation of the dosage is used. In fact, most of the published work on this subject has assumed that $p_i = F(\alpha + \beta \log(x_i))$, where F is usually either the normal or the logistic distribution. Sometimes, though, experience has suggested a transformation other than log.

In our approach, we follow the suggestion in Cox (1970, p. 110) of applying the Box-Cox transformation to the dosage level. That is to the independent variable. See Box and Tidwell (1962) for a thorough discussion and applications of transformations to independent variables. The aim in the present situation is to nearly symmetrize the original tolerance distribution, even when the assumed model is incorrect.

Thus, we tentatively assume

$$p_i(\underset{\sim}{\theta}_0) = P[T^{(\lambda_0)} < x_i^{(\lambda_0)}] = F(\alpha_0 + \beta_0 x_i^{(\lambda_0)}) \qquad (5)$$

for some parameter values α_0, β_0 and λ_0, where F is a known symmetric distribution. In fact we consider a location scale family created from a pdf $f(\cdot)$, which is itself differentiable. The vector parameter $\underset{\sim}{\theta} = (\alpha,\beta,\lambda)'$ will then be estimated by maximum likelihood. The likelihood function for $n = \sum_{i=1}^{k} n_i$ observations is given by

$$L_n(\underset{\sim}{\theta}) = \prod_{i=1}^{k} \binom{n_i}{r_i} p_i^{r_i}(\underset{\sim}{\theta})[1 - p_i(\underset{\sim}{\theta})]^{n_i - r_i}$$

so, the log-likelihood for $\underset{\sim}{\theta}$ is simply

$$\ell_n(\underset{\sim}{\theta}) = \sum_{i=1}^{k} \{\log(n_i!) - \log[(n_i - r_i)!] - \log(r_i!) + r_i \log[p_i(\underset{\sim}{\theta})] + (n_i - r_i)\log[1 - p_i(\underset{\sim}{\theta})]\}.$$

Strong consistency of the MLE is established even when the model (5) is not correct.

<u>Theorem 3</u>: For each $i = 1,\ldots,k$, let p_i be the probability of a response under the true tolerance distribution (G). Let $p_i(\underset{\sim}{\theta})$ be the probability under the transformed distribution (F) and let $\hat{p}_i$ be the observed frequency of response. If

(i) the parameter space $\Omega \subset \mathbb{R}^3$ is compact,

(ii) $\lim_{n\to\infty} n_i/n = s_i$, with $0 < s_i < 1$ $\forall$ i and $\sum_{i=1}^{k} s_i = 1$,

(iii) the function $H(\underset{\sim}{\theta}) = \sum_{i=1}^{k} s_i\left\{p_i \log\left[\frac{p_i(\underset{\sim}{\theta})}{p_i}\right] + (1 - p_i)\log\left[\frac{1 - p_i(\underset{\sim}{\theta})}{1 - p_i}\right]\right\}$ has a unique maximum at $\underset{\sim}{\theta} = \underset{\sim}{\theta}_0 = (\alpha_0,\beta_0,\lambda_0)'$;

Then, the MLE $\hat{\underset{\sim}{\theta}}_n$ is a strongly consistent estimator of $\underset{\sim}{\theta}_0$. ■

The asymptotic distribution of $\hat{\underset{\sim}{\theta}}_n$ can be *obtained based on the* asymptotic behavior of the gradient $\nabla\ell_n(\underset{\sim}{\theta})$ and Hessian $\nabla^2\ell_n(\underset{\sim}{\theta})$ of the log-likelihood. Namely, $\nabla\ell_n(\underset{\sim}{\theta})$ has components

$$\frac{\partial \ell_n(\underset{\sim}{\theta})}{\partial\theta_u} = \sum_{i=1}^{k} n_i\left\{\frac{\hat{p}_i - p_i(\underset{\sim}{\theta})}{p_i(\underset{\sim}{\theta})[1 - p_i(\underset{\sim}{\theta})]}\left(\frac{\partial p_i(\underset{\sim}{\theta})}{\partial\theta_u}\right)\right\} \quad u = 1,2,3 \qquad (6)$$

and $\nabla^2\ell_n(\underset{\sim}{\theta})$ has components

$$\frac{\partial \ell_n^2(\underset{\sim}{\theta})}{\partial\theta_u \partial\theta_v} = \sum_{i=1}^{k} n_i \left\{ \frac{\hat{p}_i[2p_i(\underset{\sim}{\theta}) - 1] - p_i^2(\underset{\sim}{\theta})}{p_i^2(\underset{\sim}{\theta})[1 - p_i(\underset{\sim}{\theta})]^2} \left(\frac{\partial p_i(\underset{\sim}{\theta})}{\partial \theta_u}\right)\left(\frac{\partial p_i(\underset{\sim}{\theta})}{\partial \theta_v}\right) + \frac{\hat{p}_i - p_i(\underset{\sim}{\theta})}{p_i(\underset{\sim}{\theta})[1 - p_i(\underset{\sim}{\theta})]} \left(\frac{\partial^2 p_i(\underset{\sim}{\theta})}{\partial\theta_u \partial\theta_v}\right)\right\}, \quad u,v = 1,2,3 \tag{7}$$

where $\frac{\partial p_i(\underset{\sim}{\theta})}{\partial\theta_u} = f(X_i)\,\frac{\partial X_i}{\partial \theta_u}$ with $X_i = \alpha + \beta\, x_i^{(\lambda)}\ \forall\, i$.

<u>Theorem 4</u>: Let the assumptions (i)-(iii) of Theorem 3 be true and suppose further that

(iv) the true parameter value $\underset{\sim}{\theta}_0$ is an interior point of Ω

(v) $\sum_{i=1}^{k} \sqrt{a_i}\, [p_i - p_i(\underset{\sim}{\theta}_0)](p_i(\underset{\sim}{\theta}_0)[1 - p_i(\underset{\sim}{\theta}_0)])^{-1}\, \nabla p_i(\underset{\sim}{\theta}_0) = \underset{\sim}{0}$

(vi) the Hessian of $H(\underset{\sim}{\theta})$, $\nabla^2 H(\underset{\sim}{\theta}) = \left(\frac{\partial^2 H(\underset{\sim}{\theta})}{\partial\theta_u \partial\theta_v}\right)_{3\times 3}$ is nonsingular at $\underset{\sim}{\theta}_0$.

Then $\sqrt{n}(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0) \xrightarrow{d} N_3(\underset{\sim}{0},\ V W V')$ as $n \to \infty$, where $V = [\nabla^2 H(\underset{\sim}{\theta}_0)]^{-1}$

and

$$W = \sum_{i=1}^{k} a_i\, p_i (p_i(\underset{\sim}{\theta}_0)[1 - p_i(\underset{\sim}{\theta}_0)])^{-1}\{1 - p_i(p_i(\underset{\sim}{\theta}_0)[1 - p_i(\underset{\sim}{\theta}_0)])^{-1}\} \cdot [\nabla p_i(\underset{\sim}{\theta}_0)][\nabla p_i(\underset{\sim}{\theta}_0)]' \tag{8}$$

with $\nabla p_i(\underset{\sim}{\theta}_0) = \left(\frac{\partial p_i(\underset{\sim}{\theta})}{\partial\theta_u}\bigg|_{\underset{\sim}{\theta}_0}\right)_{3\times 1}$. ■

5. Examples of Transformation with Probit and Logit Models

We first remark that it is perhaps computationally simplest to maximize the log-likelihood function following a two-stage procedure. That is, first fix a value of λ and maximize $\ell_n(\alpha,\beta,\lambda)$ over α and β. Then search over values of λ. A consistent estimate of the variance-covariance matrix, $\frac{1}{n} V W V'$, is obtained by replacing the true probabilities (p_i's) by the observed frequencies ($\frac{n_i}{n}$'s) and the true parameter value ($\underset{\sim}{\theta}_0$) by its MLE ($\hat{\underset{\sim}{\theta}}_n$). For illustration, we consider the data shown in Table 3. Finney (1971) analyzed these data to compare the performance of probits vs. logits.

The parameter estimates for the integrated normal model are $\hat{\alpha} = -71.1019$, $\hat{\beta} = 53.6827$, $\hat{\lambda} = -.587$, and for the logit model $\hat{\alpha} = -71.9335$, $\hat{\beta} = 36.0568$, $\hat{\lambda} = -.204$. The corresponding estimated variance-covariance matrices are

$$\begin{pmatrix} 59.5757 & & \\ -73.1211 & 89.7779 & \\ 0.5393 & -0.6625 & 0.0493 \end{pmatrix}$$

and

$$\begin{pmatrix} 86.3624 & & \\ -78.5013 & 71.3697 & \\ 0.8346 & -0.7589 & 0.0811 \end{pmatrix}.$$

Table 3

AGE OF MENARCHE IN 3918 WARSAW GIRLS
(taken from Finney (1971, p. 98))

Mean age of group (years)	No. of girls	No. having menstruated: Observed	Estimated PROBIT (original)	Estimated PROBIT ($\hat{\lambda}$=-.587)	Estimated LOGIT (original)	Estimated LOGIT ($\hat{\lambda}$=-.204)
9.21	376	0	0.10	0.00	0.76	0.20
10.21	200	0	1.08	0.24	2.06	1.07
10.58	93	0	1.25	0.50	1.74	1.10
10.83	120	2	2.81	1.51	3.34	2.37
11.08	90	2	3.53	2.37	3.72	2.91
11.33	88	5	5.51	4.38	5.36	4.56
11.58	105	10	10.05	9.01	9.33	8.51
11.83	111	17	15.56	15.15	14.19	13.69
12.08	100	16	19.70	20.20	18.06	18.16
12.33	93	29	24.72	26.10	23.15	23.89
12.58	100	39	34.51	36.86	33.26	34.78
12.83	108	51	46.64	49.80	46.27	48.46
13.08	99	47	51.69	54.73	52.46	54.60
13.33	106	67	64.78	67.76	66.67	68.70
13.58	105	81	72.95	75.28	75.42	76.87
13.83	117	88	90.00	91.69	92.79	93.69
14.08	98	79	81.36	82.20	83.51	83.72
14.33	97	90	85.65	85.62	86.97	86.78
14.58	120	113	110.61	109.99	111.45	110.90
14.83	102	95	96.89	96.07	97.05	96.45
15.08	122	117	118.26	117.16	118.00	117.23
15.53	111	107	109.01	108.05	108.55	107.88
15.58	94	92	93.06	92.35	92.61	92.09
15.83	114	112	113.39	112.70	112.87	112.32
17.58	1049	1049	1048.98	1048.59	1048.40	1047.15

Therefore, large sample 95% confidence intervals for the transformation parameter in the two situations are (-1.022, .1518) and (-.7622, .3542), neither of which covers the value $\lambda = 1$. Thus, the use of the original measurement of age is not sensible for the present situation. Finney used the original scale "as general evidence is that age itself gives a good linear relation with the response." Table 4 shows that the four series of estimated responses agree well with the observations, so there is no clear choice between the models. The chi-square statistics (without any grouping at the extremes) of the following table do show a slight preference for the probit model with age transformed by $\hat{\lambda} = -.587$.

Table 4

COMPARISON OF MODELS

Model	Estimated Median Age	X^2	Degrees of freedom
Probit (Original)	13.019	21.901	23
Probit ($\hat{\lambda}$ = -.587)	12.935	13.061	22
Logit (Original)	13.007	21.870	23
Logit ($\hat{\lambda}$ = 0.204)	12.954	17.921	22

Appendix: Proof of Theorem 4

Let us expand $\nabla \ell_n(\hat{\boldsymbol{\theta}}_n)$ in Taylor's series about $\boldsymbol{\theta}_0$, so that $\frac{1}{\sqrt{n}} \nabla \ell_n(\hat{\boldsymbol{\theta}}_n) = \frac{1}{\sqrt{n}} \nabla \ell_n(\boldsymbol{\theta}_0) + \frac{1}{n} \nabla^2 \ell_n(\boldsymbol{\theta}_{*n})[\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)]$ for some $\boldsymbol{\theta}_{*n} = \gamma_n \hat{\boldsymbol{\theta}}_n + (1-\gamma_n)\boldsymbol{\theta}_0$, $0 < \gamma_n < 1$. Since $\hat{\boldsymbol{\theta}}_n \xrightarrow{a.s.} \boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_0$ is an interior point of Ω, $\nabla \ell_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}$ for n sufficiently large, with probability one. Thus, we know that $\frac{1}{\sqrt{n}} \nabla \ell_n(\boldsymbol{\theta}_0)$ and $-\frac{1}{n} \nabla^2 \ell_n(\boldsymbol{\theta}_{*n})[\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)]$ have the same limiting distribution.

Next, let us recall that $r_i = \sum_{j=1}^{n_i} Y_j$ where $Y_j = 1$ if and only if $T_j < x_i$. Therefore, by (6) we get $\nabla \ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{k} \nabla \ell_{n_i}(\boldsymbol{\theta})$ where

$$\nabla \ell_{n_i}(\boldsymbol{\theta}) = \sum_{j=1}^{n_i} [Y_j - p_i(\boldsymbol{\theta})](p_i(\boldsymbol{\theta})[1 - p_i(\boldsymbol{\theta})])^{-1} \nabla p_i(\boldsymbol{\theta})$$

$$= \sum_{j=1}^{n_i} \mathbf{Z}_{ij}(\boldsymbol{\theta}) \quad \text{for } i = 1,\ldots,k.$$

For each i, the random vectors $\{\mathbf{Z}_{ij}(\boldsymbol{\theta}_0),\ j = 1,\ldots,n_i\}$ are iid with $E_g[\mathbf{Z}_{i1}(\boldsymbol{\theta}_0)] = \mathbf{U}_i$ and $\mathrm{Var}_g[\mathbf{Z}_{i1}(\boldsymbol{\theta}_0)] = \Psi_i$ where $\mathbf{U}_i = [\rho_i - p_i(\boldsymbol{\theta}_0)](p_i(\boldsymbol{\theta}_0)[1 - p_i(\boldsymbol{\theta}_0)])^{-1} \nabla p_i(\boldsymbol{\theta}_0)$ and $\Psi_i = \rho_i(p_i(\boldsymbol{\theta}_0)[1 - p_i(\boldsymbol{\theta}_0)])^{-1}(1 - \rho_i(p_i(\boldsymbol{\theta}_0)[1 - p_i(\boldsymbol{\theta}_0)])^{-1} \cdot [\nabla p_i(\boldsymbol{\theta}_0)][\nabla p_i(\boldsymbol{\theta}_0)]'$. Thus, applying the multivariate CLT k-times, we get $\frac{1}{\sqrt{n_i}} \nabla \ell_{n_i}(\boldsymbol{\theta}_0) \xrightarrow{d} N_3(\mathbf{U}_i, \Psi_i)$ as $n_i \to \infty$, $i = 1,\ldots,k$. Further, $\frac{1}{\sqrt{n}} \nabla \ell_n(\boldsymbol{\theta}_0) = \sum_{i=1}^{k} \sqrt{n_i/n}\ \frac{1}{\sqrt{n_i}} \nabla \ell_{n_i}(\boldsymbol{\theta}_0) \xrightarrow{d} N_3(\mathbf{0}, \sum_{i=1}^{k} s_i \Psi_i)$ in such a way that

$$\frac{1}{n} \nabla^2 \ell_n(\boldsymbol{\theta}_{*n})[\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)] \xrightarrow{d} N_3(\mathbf{0}, W) \quad \text{as } n \to \infty \tag{9}$$

with W as given in (8). Now, it can be observed that the second order partial derivatives defined by (7) are uniformly continuous for $\boldsymbol{\theta} \in \Omega$ $\forall$ u,v. Thus, since $\hat{p}_i$ is a consistent estimator of p_i, it follows that

$$\lim_{n\to\infty} \frac{1}{n} \nabla^2 \ell_n(\boldsymbol{\theta}) = \nabla^2 H(\boldsymbol{\theta}) \text{ with probability one and uniformly on } \Omega. \tag{10}$$

Next, as $\boldsymbol{\theta}_{*n} = \gamma_n \hat{\boldsymbol{\theta}}_n + (1-\gamma_n)\boldsymbol{\theta}_0$ for $0 < \gamma_n < 1$, (10) implies that

$$\lim_{n\to\infty} \frac{1}{n} \nabla^2 \ell_n(\boldsymbol{\theta}_{*n}) = \nabla^2 H(\boldsymbol{\theta}_0) \text{ in probability.}$$

Premultiplying (9) by $V = [\nabla^2 H(\boldsymbol{\theta}_0)]^{-1}$ and using Slutsky's Theorem we obtain the desired conclusion. ∎

Bibliography

Box, G. E. P. and Cox, D. R. (1964). "An analysis of transformations." J. R. Statist. Soc. B-26, 211-52.

Box, G. E. P. and Tidwell, P. W. (1962). "Transformation of the independent variables." Technometrics 4, 531-50.

Cox, D. R. (1970). The Analysis of Binary Data. Methuen, London.

Dyke, G. V. and Patterson, H. D. (1952). "Analysis of factorial arrangements when the data are proportions." Biometrics 8, 1-12.

Fienberg, S. E. (1977). The Analysis of Cross-Classified Categorical Data. The MIT Press.

Finney, D. J. (1971). Probit Analysis. Third Edition. Cambridge University Press.

Nerlove, M. and Press, S. J. (1973). "Univariate and multivariate log-linear and logistic models." R-1306-EDA/NIH, Rand Corporation, Santa Monica.

Prentice, R. L. (1976a). "Use of the logistic model in retrospective studies." Biometrics 32, 597-606.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: Technical Report # 575	2. GOVT ACCESSION NO	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): USE OF THE BOX-COX TRANSFORMATION WITH BINARY RESPONSE MODELS		5. TYPE OF REPORT & PERIOD COVERED
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Victor M. Guerrero; Richard A. Johnson		8. CONTRACT OR GRANT NUMBER(s): ONR Grant No. N00014-78-C-0722
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706		10. PROGRAM ELEMENT PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, 800 N. Quincy Street, Arlington, VA 22217		12. REPORT DATE: August 1979
		13. NUMBER OF PAGES: 21 pages
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS (of this report): Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of this document is unlimited

DISTRIBUTION STATEMENT A

Approved for public release; Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

transformations, logistic regression, dose-response

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The power transformation ~~suggested by Box and Cox (1964)~~ is applied to the odds ratio to generalize the logistic model and to parameterize a certain type of lack of fit. Transformation of the design variable within the context of the dose-response problem is also considered.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)